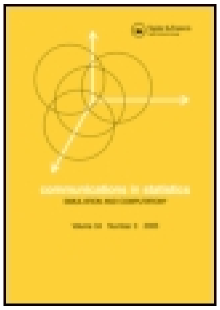

Communications in Statistics - Simulation and Computation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Determining Optimum Strata Boundaries and Sample Sizes for Skewed Population with Log-normal Distribution

M.G.M. Khan [a], Dinesh Rao [a], A. H. Ansari [b] & M. J. Ahsan [c]

[a] School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji

[b] Kejriwal Institute of Management and Development Studies, Namkum, Ranchi, India

[c] Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India

Accepted author version posted online: 07 Apr 2014.Published online: 07 Apr 2014.

To cite this article: M.G.M. Khan, Dinesh Rao, A. H. Ansari & M. J. Ahsan (2014): Determining Optimum Strata Boundaries and Sample Sizes for Skewed Population with Log-normal Distribution, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2013.819917

To link to this article: http://dx.doi.org/10.1080/03610918.2013.819917

Determining Optimum Strata Boundaries and Sample Sizes for Skewed Population with Log-normal Distribution

M.G.M. Khan and Dinesh Rao

School of Computing, Information and Mathematical Sciences

The University of the South Pacific, Suva, Fiji.

Email: khan_mg@usp.ac.fj and rao_di@usp.ac.fj

A.H. Ansari

Kejriwal Institute of Management and Development Studies

Namkum, Ranchi, India.

Email: athar_hussain1@rediffmail.com

M.J. Ahsan

Department of Statistics and Operations Research

Aligarh Muslim University, Aligarh, India.

Email: mjahsan2001@yahoo.co.in

Abstract

The method of choosing the best boundaries that make strata internally homogeneous as far as possible is known as optimum stratification. To achieve this, the strata should be constructed in such a way that the strata variances for the characteristic under study be as small as possible. If the frequency distribution of the study variable x is known, the Optimum Strata Boundaries (OSB) could be obtained by cutting the range of the distribution at suitable points. If the frequency distribution of x in unknown, it may be approximated from the past experience or some prior knowledge obtained at a recent study. Many skewed populations have Log-normal frequency distribution or may be assumed to follow approximately Log-normal frequency distribution. In this paper, the problem of finding the OSB and the optimum sample sizes within the stratum for a skewed population with

Log-normal distribution is studied. The problem of determining the OSB is redefined as the problem of determining Optimum Strata Widths (OSW) and is formulated as a Nonlinear Programming Problem (NLPP) that seeks minimization of the variance of the estimated population mean under Neyman allocation subject to the constraint that the sum of the widths of all the strata is equal to the range of the distribution. The formulated NLPP turns out to be a multistage decision problem that can be solved by dynamic programming technique. A numerical example is presented to illustrate the application and computational details of the proposed method. A comparison study is conducted to investigate the efficiency of the proposed method with other stratification methods viz Dalenius and Hodges' cum $\sqrt{f}$ method, Geometric method by Gunning and Horgan and Lavallee-Hidiroglou method using Kozak's algorithm available in the literature. The study reveals that the proposed technique is efficient in minimizing the variance of the estimate of the population mean and is useful to obtain OSB for a skewed population with Log-normal frequency distribution .

Key Words: Stratified sampling, Optimum stratification, Optimum sample size, Log-normal distribution, Nonlinear programming problem, Multistage decision problem, Dynamic programming technique.

1 Introduction

When a study variable x itself is used as a stratification variable, the problem of determining optimum strata boundaries (OSB) was first discussed by Dalenius (1950). He presented a set of minimal equations whose solution could provide the OSB. Unfortunately, the exact solution of these equations could not usually be obtained because of their implicit nature. Several attempts have been made by many authors including Dalenius and Gurney (1951), Mahalanobis (1952), Hansen, *et al.* (1953), Aoyama (1954), Ekman (1959), Dalenius and Hodges (1959), Sethi (1963), Serfling (1968), and Singh (1975) for choosing the OSB. These authors proposed the methods that give approximate strata boundaries by using classical approach.

Many authors such as Unnithan (1978), Lavallée and Hidiroglou (1988), Hidiroglou and Srinath (1993), Sweet and Sigman (1995) and Rivest (2002) suggested some iterative procedures to determine OSB. These algorithms require an initial approximate solution to start with. Also there is no guarantee that the algorithm will provide the global minimum in the absence of a suitable approximate initial solution and the variance function have more than one local minima. Moreover, the convergence of some of these algorithms are slow or non-existent (see Detlefsen and Veum 1991 and Khan *et al.* 2008).

Gunning and Horgan (2004) developed an approximate method of stratification for positively skewed populations. They showed that their algorithm is much easier and more efficient than the cum $\sqrt{f}$ method of Dalenius and Hodges (1959) and Lavallee-Hidiroglou (1988) method.

Niemiro (1999) proposed a random search method for optimum stratification but the algorithm did not guarantee that it leads to global optimum and also goes wrong in case of a large population, as it requires too many iteration steps. Lednicki and Wieczorkowski (2003) presented a method of stratification based on Rivest (2002) using the simplex method of Nelder and Mead (1965) but the method was rather slow and may not provide the best solution in the case of large number of variables. Later Kozak (2004) presented the modified random search algorithm as a method of the optimal stratification; as a random search, it does not guarantee reaching the global optimum (Kozak 2004). This algorithm was later found very efficient in stratification (e.g., Baillargeon and Rivest 2009).

Another method of stratification that has been proposed in the literature is due to Bühler and

ACCEPTED MANUSCRIPT

Deutler (1975). They formulated the problem of determining OSB as an optimization problem and developed a computational technique to solve the problem using dynamic programming. A brief review of this method can also be found in Khan *et al.*(2008). Later the technique was extended by Lavallée (1987) and Lavallée (1988) for two-way stratification. Khan *et al.*(2002, 2005, 2008), and Nand and Khan (2008) also extended this procedure for determining OSB for the study variables with different frequency functions. They considered the problem of finding OSB as an equivalent problem of determining Optimum Strata Width (OSW), which is formulated as a Nonlinear Programming Problem (NLPP) and solved by dynamic programming technique. The advantage of this technique is that it gives the optimum solution of the objective function and it does not require an initial solution, if the frequency distribution of the study variable is known and the number of strata is fixed in advance.

In this paper, a technique using the dynamic programming approach is developed to determine the OSB and the optimum sample size for each stratum under Neyman allocation for a positively skewed population with Log-normal distribution as in practice many populations have Log-normal distribution or can be assumed to have approximately Log-normal distribution. Section 2 provides the detailed formulation of the problem of finding OSW as an NLPP. The solution procedure to solve the NLPP is then discussed in Section 3 and the computational details of the technique is illustrated through a numerical example in Section 4. Finally, in Section 5, a comparison study is carried out to investigate the effectiveness of the proposed method with Dalenius and Hodges' cum $\sqrt{f}$ method, Gunning and Horgan's Geometric method, Lavallee-Hidiroglou's method using Kozak (2004) algorithm that are available in the literature as in Baillargeon and Rivest (2009).

ACCEPTED MANUSCRIPT

2 The Formulation

Let X be a random study variable with probability density function $f(x)$, $a \leq x \leq b$. To estimate the population mean μ by a stratified sample, the range of X is partitioned into L strata $[a, x_1], (x_1, x_2]$, $\ldots, (x_{L-1}, b]$ such that

$$a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b. \tag{1}$$

Suppose that from stratum h ($h = 1, 2, \ldots, L$), which contains N_h units, a simple random sample of size n_h is selected. Let y_{hj} denote the value of the $j^{th}(j = 1, 2, \ldots, n_h)$ unit in the h^{th} stratum. Then the stratified sample mean $\overline{x}_{st} = \sum_{h=1}^{L} W_h \overline{x}_h$ will be an unbiased estimate of μ with variance

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h \sigma_h^2 \left(\frac{W_h}{n_h} - \frac{1}{N} \right), \tag{2}$$

where $W_h = N_h/N$, $\overline{x}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$, $\sigma_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left(y_{hj} - \mu_h \right)^2$ and $\mu_h = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$.

When the frequency function $f(x)$ is known, the values of W_h and σ_h^2 in (2) can be obtained by

$$W_h = \int_{x_{h-1}}^{x_h} f(x) dx, \tag{3}$$

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x) dx - \mu_h^2, \tag{4}$$

$$\text{where} \quad \mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x) dx \tag{5}$$

is the mean and (x_{h-1}, x_h) are the boundaries of h^{th} stratum.

Using the above values of W_h, μ_h and σ_h^2, the RHS of (2) can be expressed as a function of x_h and n_h, that is,

$$V(\overline{x}_{st}) = V(\overline{x}_{st} | x_1, \ldots, x_{L-1}, n_1, \ldots, n_L).$$

Further, if population mean is estimated with a fixed total sample size:

$$n = \sum_{h=1}^{L} n_h,$$

then under Neyman allocation, n_h; $(h = 1, 2, ..., L)$ are given by:

$$n_h = n \cdot \frac{W_h \sigma_h}{\sum_{h=1}^{L} W_h \sigma_h}. \tag{6}$$

If n_h; $(h = 1, 2, ..., L)$ are fixed under Neyman allocation, the objective of the optimum stratification is to determine the stratum boundary points $x_1, \ldots, x_{L-1}$ such that $V(\overline{x}_{st})$ is minimum subject to the restrictions that

$$2 \leq n_h \leq N_h. \tag{7}$$

The restrictions $n_h \leq N_h$ are imposed to avoid the over sampling, which may be the case, especially, when the population is skewed. Whereas, the restrictions $2 \leq n_h$ are imposed, when the stratum variances σ_h^2 are needed to be estimated (see Khan *et al*. 1997, 2003).

From (2), it can be seen that the second term does not have any influence on the sample size as it is independent of n_h. Thus, omitting the term and substituting (6), the variance $V(\overline{x}_{st})$ in (2) is reduced to:

$$V(\overline{x}_{st}) \doteq \frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{n}. \tag{8}$$

However, for a fixed total sample size n, the minimization of (8) is equivalent to minimizing (see Khan et al., 2005):

$$\sum_{h=1}^{L} W_h \sigma_h. \tag{9}$$

Thus the problem of determining OSB and the optimum sample size may be stated as:

$$\text{Minimize} \quad \left\{ \sum_{h=1}^{L} W_h \sigma_h | a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b; 2 \leq n_h \leq N_h \right\}. \tag{10}$$

Further, from (6),(7) and (9), it can be seen that the restrictions $n_h \leq N_h$ are satisfied, if the following conditions hold:

$$\sigma_h \leq \frac{\sum_{h=1}^{L} W_h \sigma_h}{n/N}.$$

Similarly, the restrictions $2 \leq n_h$ are satisfied, if

$$W_h \sigma_h \geq \frac{2 \sum_{h=1}^{L} W_h \sigma_h}{n}.$$

Let $f(x)$ be the frequency function and x_0 and x_L are the smallest and largest values of x. If the population mean is estimated under (6), then the problem of determining the strata boundaries is equivalent to cut up the range,

$$d = x_L - x_0, \tag{11}$$

at intermediate points $x_1 \leq x_2 \leq, ..., \leq x_{L-1}$ such that $\sum_{h=1}^{L} W_h \sigma_h$ in (10) is minimum.

If $f(x)$ is integrable, using the expressions (3), (4) and (5), W_h, σ_h^2 and μ_h are obtained as a function of the boundary points x_h and x_{h-1}. Thus the objective function in (10) could be expressed as a function of boundary points x_h and x_{h-1}, that is

$$\phi_h(x_{h-1}, x_h) = W_h \sigma_h.$$

Thus, the problem (10) can be treated as an optimization problem to find x_1, x_2, ..., x_{L-1} to:

$$\begin{aligned} &\text{Minimize} \quad \sum_{h=1}^{L} \phi_h(x_{h-1}, x_h), \\ &\text{subject to} \quad a = x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L = b. \end{aligned} \tag{12}$$

Bühler and Deutler (1975) have suggested a recursive optimization method for solving (12) using a dynamic programming technique (also see Khan *et al.* 2008).

Khan *et al.* (2002, 2005, 2008), and Nand and Khan (2008) treated the problem (12) as an equivalent problem of determining OSW as follows:

Let $y_h = x_h - x_{h-1} \geq 0$ be the width of the h^{th} $(h = 1, 2, ..., L)$ stratum.

With the above definition of y_h, the range of the distribution given in (11) may be expressed as the function of the stratum widths as:

$$\sum_{h=1}^{L} y_h = \sum_{h=1}^{L} (x_h - x_{h-1}) = x_L - x_0 = d. \tag{13}$$

The k^{th} stratification point x_k; $(k = 1, 2, ..., L - 1)$ is then expressed as:

$$\begin{aligned} x_k &= x_0 + y_1 + y_2 + ... + y_k \\ &= x_{k-1} + y_k, \end{aligned}$$

which is a function of k^{th} stratum width and $(k - 1)^{th}$ stratum boundary.

Adding (13) as a constraint, the problem (12) can be treated as an equivalent problem of determining OSW as:

$$\begin{aligned} \text{Minimize} \quad & \sum_{h=1}^{L} \phi_h(y_h, x_{h-1}), \\ \text{subject to} \quad & \sum_{h=1}^{L} y_h = d, \\ \text{and} \quad & y_h \geq 0; h = 1, 2, ..., L. \end{aligned} \tag{14}$$

Initially, x_0 is known. Therefore, the first term, that is, $\phi_1(y_1, x_0)$ in the objective function of NLPP (14) is a function of y_1 alone. Once y_1 is known, the next stratification point $x_1 = x_0 + y_1$ will be known and the second term in the objective function $\phi_2(y_2, x_1)$ will become a function of y_2 alone. Thus, stating the objective function as a function of y_h alone, we may rewrite the NLPP (14) as:

$$\begin{aligned} \text{Minimize} \quad & \sum_{h=1}^{L} \phi_h(y_h), \\ \text{subject to} \quad & \sum_{h=1}^{L} y_h = d, \\ \text{and} \quad & y_h \geq 0; \quad h = 1, 2, ..., L. \end{aligned} \tag{15}$$

When the study variable has a Log-normal frequency function, the formulation of the problem of determining OSW may be expressed as an NLPP as discussed in Section 2.1 below.

2.1 The Problem of OSB for Skewed Population with Log-normal Distribution

The Log-normal distribution is a positively skewed distribution, meaning that most of the distribution is concentrated around the left end. Surveyors may use the Log-normal distribution for a positive valued study variable that might increase without limit, such as the value of securities in financial problem or the value of properties in real estate or the failure rate of electronic parts in engineering problem.

A variable X is Log-normally distributed if $Y = ln(X)$ is normally distributed where "ln" stands for the natural logarithm. The general formula for the probability density function of the Log-normal distribution is

$$f(x) = \frac{\exp\left[-\left(((\ln(x) - \mu)/m))^2 / \left(2\sigma^2\right)\right)\right]}{x\sigma\sqrt{2\pi}}; \qquad x > 0, \quad \mu \in \mathbb{R}, \quad m > 0, \quad \sigma > 0, \tag{16}$$

where σ is the shape parameter, μ is the location parameter and m is the scale parameter.

With $m = 1$,(16) gives the Log-normal density as

$$f(x) = \frac{\exp\left[-\left((\ln(x)-\mu)^2 / \left(2\sigma^2\right)\right)\right]}{x\sigma\sqrt{2\pi}}; \qquad x > 0, \quad \mu \in \mathbb{R}, \quad \sigma > 0. \tag{17}$$

Using the definitions (3), (5), (4) and (17), the terms W_h, μ_h and σ_h^2 can be expressed as

$$W_h = \frac{1}{2}\left(erf\left(\frac{\ln(y_h + x_{h-1}) - \mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - \mu}{\sigma\sqrt{2}}\right)\right), \tag{18}$$

$$\mu_h = \exp\left(\frac{\sigma^2}{2} + \mu\right) \frac{erf\left(\frac{\ln(y_h + x_{h-1}) - \mu - \sigma^2}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - \mu - \sigma^2}{\sigma\sqrt{2}}\right)}{erf\left(\frac{\ln(y_h + x_{h-1}) - \mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - \mu}{\sigma\sqrt{2}}\right)}, \tag{19}$$

$$\begin{aligned}\sigma_h^2 = {} & \frac{1}{\left[erf\left(\frac{\ln(y_h + x_{h-1}) - \mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - \mu}{\sigma\sqrt{2}}\right)\right]^2} \\ & \left\{\left[\exp\left(2\sigma^2 + 2\mu\right)\left(erf\left(\frac{\ln(y_h + x_{h-1}) - 2\sigma^2 - \mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - 2\sigma^2 - \mu}{\sigma\sqrt{2}}\right)\right)\right]\right. \\ & \left[erf\left(\frac{\ln(y_h + x_{h-1}) - \mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - \mu}{\sigma\sqrt{2}}\right)\right] - \left[\exp\left(\frac{\sigma^2}{2} + \mu\right)\right. \\ & \left.\left.\left(erf\left(\frac{\ln(y_h + x_{h-1}) - \sigma^2 - \mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{\ln(x_{h-1}) - \sigma^2 - \mu}{\sigma\sqrt{2}}\right)\right)\right]^2\right\}.\end{aligned} \tag{20}$$

Note that an error function (erf) is used to counter the integrations with Log-normal density function. The error function is defined as

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt. \tag{21}$$

The probability that a Log-normal variate assumes a value in the range $[z_1, z_2]$ is given by:

$$\frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} \exp\left(\frac{-x^2}{2}\right) dx = \frac{1}{2}\left[erf(z_2) - erf(z_1)\right]. \tag{22}$$

Common properties of the error functions include:

$$erf(-z) = -erf(z),\ erf(0) = 0,\ erf(\infty) = 1,\ erf(-\infty) = -1. \tag{23}$$

Using (18), (20) and (21) the NLPP (15) may be expressed as:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{h=1}^{L} \left\{ \begin{array}{l} \frac{1}{2} Sqrt \left\{ \left[\exp\left(2\sigma^2 + 2\mu\right) \left(erf\left(\frac{\ln(y_h + x_{h-1}) - 2\sigma^2 - \mu}{\sigma\sqrt{2}} \right) \right.\right.\right. \\
\qquad\qquad \left.\left. - erf\left(\frac{\ln(x_{h-1}) - 2\sigma^2 - \mu}{\sigma\sqrt{2}} \right)\right)\right] \\
\qquad\qquad \left[erf\left(\frac{\ln(y_h + x_{h-1}) - \mu}{\sigma\sqrt{2}} \right) - erf\left(\frac{\ln(x_{h-1}) - \mu}{\sigma\sqrt{2}} \right)\right] \\
\qquad\qquad - \left[\exp\left(\frac{\sigma^2}{2} + \mu \right) \left(erf\left(\frac{\ln(y_h + x_{h-1}) - \sigma^2 - \mu}{\sigma\sqrt{2}} \right) \right.\right. \\
\qquad\qquad \left.\left.\left. - erf\left(\frac{\ln(x_{h-1}) - \sigma^2 - \mu}{\sigma\sqrt{2}} \right)\right)\right]^2 \right\}
\end{array} \right\}, \\
\text{subject to} \quad & \sum_{h=1}^{L} y_h = d, \\
\text{and} \quad & y_h \geq 0; \quad h = 1, 2, ..., L.
\end{aligned} \tag{24}$$

Treating (24) as a multistage decision problem, the NLPP may be solved for determining the OSW using the dynamic programming technique. At each stage the value of the OSW and hence the OSB for a stratum as well as its optimum sample size is worked out with a forward recursive equation as discussed in Section 3.

Note that upon determining the optimum boundary points (x_{h-1}, x_h) of the hth stratum, the problem of determining its optimum sample size, n_h, can be solved by using (3) - (6).

3 The Solution using Dynamic Programming Technique

The NLPP (24) is a multistage decision problem in which the objective function and the constraints are separable functions of y_h, which allow us to use a dynamic programming technique. A solution procedure using such a dynamic programming technique is discussed in Khan *et al.* (2008), which is summarized below:

Consider a subproblem of (24) of first $k(< L)$ strata, that is:

$$\begin{aligned} \text{Minimize} \quad & \sum_{h=1}^{k} \phi_h(y_h), \\ \text{subject to} \quad & \sum_{h=1}^{k} y_h = d_k, \\ \text{and} \quad & y_h \geq 0; \quad h = 1, 2, ..., k. \end{aligned} \tag{25}$$

where $d_k < d$ is the total width available for division into k strata or the state value at stage k. Note that $d_k = d$ for $k = L$.

Using the Bellman's (1957) principle of optimality, we get the recursive relation of dynamic programming technique as:

$$\Phi_k(d_k) = \min_{0 \leq y_k \leq d_k} \left[\phi_k(y_k) + \Phi_{k-1}(d_k - y_k) \right], \quad k \geq 2. \tag{26}$$

For the first stage, that is, for $k = 1$:

$$\Phi_1(d_1) = \phi_1(d_1) \implies y_1^* = d_1, \tag{27}$$

where $y_1^* = d_1$ is the optimum width of the first stratum. The relations (26) and (27) are solved recursively for each $k = 1, 2, ..., L$ and $0 \leq d_k \leq d$, and $\Phi_L(d)$ is obtained. From $\Phi_L(d)$ the optimum width of L^{th} stratum, y_L^*, is obtained. From $\Phi_{L-1}(d - y_L^*)$ the optimum width of $(L-1)^{th}$ stratum, y_{L-1}^*, is obtained and so on until y_1^* is obtained. The details of the solution procedure can be seen

ACCEPTED MANUSCRIPT

in Khan *et al.* (2008).

We also define $\Phi_k(d_k) = 0$ for $k = 0$ and $\Phi_k(d_k) = \infty$ if

$$\sigma_k > \frac{\Phi_L(d)}{n/N} \quad or \quad W_k\sigma_k < \frac{2\Phi_L(d)}{n}$$

This takes care of the restrictions $2 \leq n_h \leq N_h$ given in (7) while solving the recursive equations (26) and (27) for the optimum stratum widths y_k; $(k = 1, 2, ..., L)$ using the proposed technique (see Khan *et al*. 1997, 2003).

4 Numerical Illustration

In this section the computational details of the solution procedure developed in Section 3 for the NLPP (24) is presented.

Assume that x follows the standard Log-normal distribution in the interval [0.00001, 13.00001], that is, $a = x_0 = 0.00001$, $b = x_L = 13.00001$, $\mu = 0$ and $\sigma = 1$. This implies that $d = x_L - x_0 = 13$. Then the NLPP (24) is expressed as:

ACCEPTED MANUSCRIPT

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{h=1}^{L} \left\{ \begin{array}{l} \frac{1}{2} Sqrt \left\{ \left[\exp(2) \left(erf\left(\frac{\ln(y_h + x_{h-1}) - 2}{\sqrt{2}} \right) \right. \right. \right. \\ \qquad \left. \left. - erf\left(\frac{\ln(x_{h-1}) - 2}{\sqrt{2}} \right) \right) \right] \\ \qquad \left[erf\left(\frac{\ln(y_h + x_{h-1})}{\sqrt{2}} \right) - erf\left(\frac{\ln(x_{h-1})}{\sqrt{2}} \right) \right] \\ \qquad - \left[\exp\left(\frac{1}{2}\right) \left(erf\left(\frac{\ln(y_h + x_{h-1}) - 1}{\sqrt{2}} \right) \right. \right. \\ \qquad \left. \left. \left. - erf\left(\frac{\ln(x_{h-1}) - 1}{\sqrt{2}} \right) \right) \right]^2 \right\} \end{array} \right\}, \\
\text{subject to} \quad & \sum_{h=1}^{L} y_h = 13, \\
\text{and} \quad & y_h \geq 0; \quad h = 1, 2, ..., L. \qquad (28)
\end{aligned}
$$

Also

$$
\begin{aligned}
x_{k-1} &= x_0 + y_1 + y_2 + ... + y_{k-1} \\
&= 0.00001 + y_1 + y_2 + ... + y_{k-1} \\
&= d_{k-1} + 0.00001 \\
&= d_k - y_k + 0.00001.
\end{aligned}
$$

Substituting this value of x_{k-1} in (28) and using (27) and (26), the recurrence relations for solving NLPP (28) are obtained as:

For first stage ($k = 1$):

$$
\begin{aligned}
\Phi_1(d_1) \; &= \tfrac{1}{2} Sqrt \left\{ \left[\exp(2) \left(erf\left(\tfrac{\ln(d_1 + 0.00001) - 2}{\sqrt{2}} \right) - erf\left(\tfrac{\ln(0.00001) - 2}{\sqrt{2}} \right) \right) \right] \left[erf\left(\tfrac{\ln(d_1 + 0.00001)}{\sqrt{2}} \right) \right. \right. \\
&\left. \left. - erf\left(\tfrac{\ln(0.00001)}{\sqrt{2}} \right) \right] - \left[\exp\left(\tfrac{1}{2}\right) \left(erf\left(\tfrac{\ln(d_1 + 0.00001) - 1}{\sqrt{2}} \right) - erf\left(\tfrac{\ln(0.00001) - 1}{\sqrt{2}} \right) \right) \right]^2 \right\} \qquad (29)
\end{aligned}
$$

at $y_1 = d_1$,

and for the stages $k \geq 2$:

$$\Phi_k(d_k) = \min_{0 \leq y_k \leq d_k} \left\{ \begin{array}{l} \frac{1}{2} Sqrt \left\{ \left[\exp(2) \left(erf\left(\frac{\ln(d_k + 0.00001) - 2}{\sqrt{2}} \right) \right. \right. \right. \\ \qquad \left. \left. - erf\left(\frac{\ln(d_k - y_k + 0.00001) - 2}{\sqrt{2}} \right) \right) \right] \\ \qquad \left[erf\left(\frac{\ln(d_k + 0.00001)}{\sqrt{2}} \right) - erf\left(\frac{\ln(d_k - y_k + 0.00001)}{\sqrt{2}} \right) \right] \\ \qquad - \left[\exp\left(\frac{1}{2}\right) \left(erf\left(\frac{\ln(d_k + 0.00001) - 1}{\sqrt{2}} \right) \right. \right. \\ \qquad \left. \left. \left. - erf\left(\frac{\ln(d_k - y_k + 0.00001) - 1}{\sqrt{2}} \right) \right) \right]^2 \right\} \\ \qquad + \Phi_{k-1}(d_k - y_k) \end{array} \right\}. \tag{30}$$

Solving the recursive equations (29) and (30) by executing a computer program developed for the solution procedure described in Section 3, the OSWs are obtained. The results of optimum strata widths y_h^* and hence the optimum strata boundaries $x_h^* = x_{h-1}^* + y_h^*$ along with the values of the objective function $\sum_{h=1}^{L} \phi_h(y_h)$ for $L = 2, 3, 4, 5$ and 6 are presented in Table 1. The table also presents the sample sizes (n_h; $h = 1, 2, ..., L$) using (3) - (6) for a fixed total sample size $n = 100$.

5 Comparison Study

In this section, a comparison study is carried out to compare and investigate the effectiveness of the proposed dynamic programming method with the other methods available in the literature. The study is undertaken to compare the following methods:

1. Dalenius and Hodges' cum $\sqrt{f}$ (1959) method.

2. Geometric method by Gunning and Horgan (2004).

3. Generalized Lavallee-Hidiroglou (1988) method using Kozak's (2004) algorithm.

For the purpose of comparison, ten artificial skewed populations that follow Log-normal distribution were randomly generated using the **R** software for various combinations of parameters, such as the shape parameter (σ) that varies from 0.2 to 1.2, skewness that varies from 0.6 to 6.6 and population size (N) that varies from 1000 to 15000, etc.

For these populations the OSB are determined by using the proposed dynamic programming method as discussed in previous sections. For each population the stratification is made for 5 different number of strata, i.e. $L = 2, 3, 4, 5$ and 6. The variance

$$V(\overline{x}_{st})^{*} = \sum_{h=1}^{L} W_h \sigma_h^2 \left(\frac{W_h}{n_h} - \frac{1}{N} \right),$$

is calculated, which is used to compare the efficiency of the different methods. For each method, the OSB obtained along with the result of $V(\overline{x}_{st})^{*}$, stratum size (N_h), optimum sample size (n_h) with a fixed n that varies from 100 to 1500 are presented in Tables 2-11 in the Appendix. In last column of Table 2, the stratum variance σ_h^2 are also presented for the proposed method. The minimum value (x_0) and the range of the distribution (d) required to determine the OSB of each population are different, which are captioned in each table.

The results for the proposed method are obtained by solving the recursive equations (27) and (28) using a computer program coded in c++. Whereas, the results for other methods are obtained by using the **R** package "stratification", version 2.2-3, developed by Baillargeon and Rivest (2009, 2011) to undertake the comparison.

In comparison of the proposed method with cum $\sqrt{f}$ and Geometric methods, it has been observed that the proposed method provides least variance of the estimate (i.e. $V(\overline{x}_{st})^{*}$) in almost all the cases. The study also reveals that the proposed method performs even better than the two methods when the skewness increases. The Geometric method performs very badly as compared to others and may not be useful as it violates the required restrictions on sample sizes given in (7), especially, when L increases. The results support the findings of Kozak and Verma (2006) which showed that the Geometric method is less efficient than L-H method. However, the findings in this study contradict with that of Gunning and Horgan (2004), which showed that the geometric method is more efficient than the cum $\sqrt{f}$ method. On the other hand, although, the cum $\sqrt{f}$ method performs better than the Geometric method, it sometimes fails to determine the OSB, if L is large and nclass is small (e.g. see Table 9 for $L = 6$).

Whereas, the comparison between L-H method and other methods reveals that L-H method provides least variances in all cases. However, the performance of proposed method is very similar to the L-H method as there is not much significant difference in the variances. It can be noted that the comparison is made by using the criteria of minimum variance calculated using all the data values that fall within the stratum boundaries from the dataset of a population. Except the proposed method, the minimum variances are calculated for each method using the OSBs that are obtained using the dataset which is the basis this comparison. Whereas, the minimum variances are calculated for the proposed method using the OSBs that are obtained using the values that fall on the density function of the log-normal distributions and not using the values in the dataset. Because of the difference in the procedure used, L-H method produces slight better results over the proposed method. Further, an advantage of the proposed method over L-H method is that it needs neither any initial solution nor the complete dataset. In many situations, the complete dataset may not be available, in such cases the proposed method works as it requires only the parameters of the distribution.

6 Summary

This paper deals with the problem of determining optimum strata boundaries (OSB) and the sample allocation to strata for a skewed population that has Log-normal frequency distribution. The problem is formulated as an NLPP, which is solved by developing a method using dynamic programming technique.

A numerical example on determining OSB is presented to show the computational details and the applications of proposed method using dynamic programming technique. Based on the results, we conclude that the proposed method is helpful in choosing the best boundary points for stratification. Furthermore, a comparison study is carried out using ten artificial populations to compare efficiency of the proposed method with the cum $\sqrt{f}$, Geometric and L-H methods. The results in the study reveal that the proposed method and the L-H method are more efficient than the cum $\sqrt{f}$ and Geometric methods in minimizing the variance of the estimate of the population mean.

The basic advantage of dynamic programming over the classical optimization techniques is that it can determine OSB efficiently, when the density function of the population is known or approximately known from the previous studies. Many other iterative methods are also available for determining strata boundaries but these iterative methods require approximate initial solutions. Also there is no guarantee that an iterative method will converge and give the global minimum variance in the absence of a suitably chosen initial solution (Amini *et al.*, 1990; Hillier and Lieberman, 2010; Khan, *et al.*, 2008). Whereas, the proposed method does not require any initial approximate solution.

More importantly, the proposed technique has a wide scope of application as compared to other methods. In practice, the complete dataset of the study variable is unknown, which diminishes the

ACCEPTED MANUSCRIPT

uses of many stratification techniques. In such a situation, only the proposed technique can be used as it requires only the values of parameters of the population which can easily be available from the past studies. Thus, we may conclude that the proposed method is relatively efficient and may be useful for determining the OSB for any skewed population.

Acknowledgement

The authors are grateful to Prof. N. Balakrishnan, the Editor in Chief, and the referees for their invaluable comments and suggestions to improve the manuscript.

References

[1] Amini, A.A., Weymouth, T.E., and Jain, R.C. (1990). Using Dynamic Programming for Solving Variational Problems in Vision.*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9), 855-867.

[2] Aoyama, H. (1954). A study of stratified random sampling. *Annals of the Institute of Statistical Mathematics*, 6, 1-36.

[3] Baillargeon, S., and Rivest, L.-P. (2009). A General Algorithm for Univariate Stratification. *International Statistical Review*, 77(3) 331-344.

[4] Baillargeon, S., and Rivest, L.-P. (2011). The construction of stratified designs in R with package stratification. *Survey Methodology*, 37(1), 53-65.

[5] Bellman, R.E. (1957). *Dynamic Programming*. Princetown University Press, New Jersey.

[6] Bühler, W., and Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, Vol. 22, pp. 161-175.

[7] Dalenius, T. (1950). The problem of optimum stratification-II. *Skand. Aktuartidskr*, 33, 203-213.

[8] Dalenius, T., and Gurney, M. (1951). The problem of optimum stratification. *Skand. Aktuartidskr*, 34, 133-148.

[9] Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

[10] Detlefsen, R.E., and Veum, C.S. (1991). Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section*, ASA, pp. 214-219.

[11] Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Annals of the Institute of Statistical Mathematics*, 30, 1131-1134.

[12] Gunning P, Horgan JM. (2004) A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 30(2), 159-166.

[13] Hansen, M.H., and Hurwitz, W.N. (1953). On the Theory of Sampling from Finite Population. *Ann. Math. Statist.*, 14, 333-362.

[14] Hidiroglou, M.A., and Srinath, K.P. (1993). Problems Associated with Designing Subannual Bussiness Surveys. *Journal of Bussiness and Economic Statistics*, 11, 397-405.

[15] Hillier, F.S., and Lieberman, G.J. (2010). *Introduction to Operations Research*. McGraw-Hill, New York.

[16] Khan, M.G.M.; Ahsan, M. J. and Jahan, N. (1997). Compromise allocation in multivariate stratified sampling: an integer solution. *Naval Research Logistics*, 44, 69-79.

[17] Khan, E.A., Khan, M.G.M., and Ahsan, M.J. (2002). Optimum stratification: A mathematical programming approach. *Culcutta Statistical Association Bulletin*, 52 (special), 205-208.

[18] Khan, M.G.M., Khan, E.A. and Ahsan, M. J. (2003). An optimal multivariate stratified sampling design using dynamic programming. *Australian and New Zealand Journal of Statistics*, 45(1), 107-113.

[19] Khan, M.G.M., Najmussehar, and Ahsan, M.J. (2005). Optimum stratification for exponential study variable under Neyman allocation. *Journal of Indian Society of Agricultural Statistics*, 59(2), 146-150.

[20] Khan, M.G.M., Nand, N., and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34(2), 205-214.

[21] Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

[22] Kozak, M. and Verma, M. (2006). Geometric versus optimization approach to stratification: a comparison of efficiency. *Survey Methodology*, 32(2), 157-163.

[23] Lavallée, P. (1987). Some contributions to optimal stratification. Master Thesis, Carleton University, Ottawa, Canada.

[24] Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Virginia, 646-651.

[25] Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

[26] Lednicki, B., and Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

[27] Mahalanobis, P.C. (1952). Some aspects of the design of sample surveys. *Sankhya*, 12, 1-7.

[28] Nand, N., and Khan, M.G.M. (2008). Optimum stratification for Cauchy and Power type study variables. *Journal of Applied Statistical Science*, 16(4), 65-74.

[29] Niemiro, W. (1999). Konstrukcja optymalnej stratyfikacja metoda poszukiwan losowych. (Optimal Stratification using Random Search Method). *Wiadomosci Statystyczne*, 10, 1-9.

[30] Rivest, L.P. (2002). A generalization of Lavallée and Hidiroglou algorithm for stratification in business survey. *Survey Methodology*, 28, 191-198.

[31] Serfling, R. J. (1968). Approximately optimal stratification. *Journal of American Statistical Association.*, 63, 1298-1309.

[32] Sethi, V.K. (1963). A note on optimum stratification of population for estimating the population mean. *Australian Journal of Statistics*, 5, 20-33.

[33] Singh, R. (1975). On optimum stratification for proportional allocation. *Sankhya*, 37, C, 109-115.

[34] Sweet, E.M., and Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data, *Proceedings of the Survey Research Methods Section*, American Statistical Association, 491-496.

[35] Unnithan, V.K.G. (1978). The minimum variance boundry points of stratification. *Sankhya*, 40, C, 60-72.

Table 1: OSW, OSB, Sample Size and the value of objective function for standard Log-normal study variable.

No. of strata L	OSW (y_h^*)	OSB $(x_h^* = x_{h-1}^* + y_h^*)$	Optimum sample sizes $n_h = n \cdot \frac{W_h\sigma_h}{\sum_{h=1}^{L} W_h\sigma_h}$	Objective function $\sum_{h=1}^{L} \phi_h(y_h) = \sum_{h=1}^{L} W_h\sigma_h$
2	y_1^* = 2.23652	x_1^*=2.23653	51	0.8569355124
	y_2^* = 10.76348		49	
	y_1^* = 1.30859	x_1^*=1.30860	35	
3	y_2^* = 2.35085	x_2^*= 3.65945	32	0.5773613579
	y_3^* = 9.34056		33	
	y_1^* = 0.95459	x_1^*=0.95460	26	
4	y_2^* = 1.25278	x_2^*= 2.20738	25	0.4358095763
	y_3^* = 2.53417	x_3^*= 4.74155	24	
	y_4^* = 8.25846		25	
	y_1^* = 0.76589	x_1^*=0.76590	20	
	y_2^* = 0.84332	x_2^*=1.60922	20	
5	y_3^* = 1.36367	x_3^*= 2.97289	20	0.3501356776
	y_4^* = 2.62141	x_4^*= 5.59430	20	
	y_5^* = 7.40571		20	
	y_1^* = 0.64767	x_1^*= 0.64768	17	
	y_2^* = 0.63431	x_2^*= 1.28199	17	
6	y_3^* = 0.90957	x_3^*= 2.19156	16	0.2926636591
	y_4^* = 1.44256	x_4^*= 3.63412	16	
	y_5^* = 2.65047	x_5^*= 6.28459	16	
	y_6^* = 6.71542		18	

Appendix

Table 2: OSB and Optimum sample sizes for skewness = 0.5994, μ = 0.00009935132, σ = 0.1975361, N = 15000, n = 1000, nclass = 50, x_0 = 0.49410530 and d = 1.58890220.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.				
	OSB	$V(\overline{x}_{st})^*$	N_h	n_h	OSB	$V(\overline{x}_{st})^*$	N_h	n_h	OSB	$V(\overline{x}_{st})^*$	N_h	n_h	OSB	$V(\overline{x}_{st})^*$	N_h	n_h	σ_h^2
2	1.03	.000014	8493	483	1.01	.000014	7924	431	1.04	.000014	8566	490	1.04	.000014	8657	498	.01115
			6507	517			7076	569			6434	510			6343	502	.02102
3	0.91	.000007	4660	280	0.80	.000013	1896	64	0.94	.000007	5517	358	0.94	.000007	5506	357	.00701
	1.16		7013	383	1.29		11601	834	1.16		6155	302	1.16		6176	304	.00405
			3327	337			1503	102			3328	340			3318	339	.01733
4	0.84	.000004	2905	191	0.71	.000008	576	19	0.87	.000004	3620	259	0.87	.000004	3716	268	.00518
	1.03		5588	299	1.01		7348	428	1.03		4807	221	1.04		4937	233	.00222
	1.26		4670	288	1.45		6618	518	1.23		4302	234	1.24		4231	235	.00306
			1837	222			458	35			2271	286			2116	264	.01552
5	0.81	.000003	2193	164	0.66	.000006	267	9	0.84	.000003	2734	220	0.84	.000003	2730	219	.00425
	0.97		4386	240	0.88		3560	171	0.97		3877	184	0.97		3909	187	.00151
	1.13		4433	243	1.17		8027	555	1.11		3897	187	1.11		3931	191	.00155
	1.32		2782	181	1.56		2977	250	1.29		2961	183	1.29		2961	187	.00261
			1206	172			169	15			1531	226			1469	216	.01426
6	0.78	.000002	1555	128	0.63	.000004	148	5	0.81	.000002	2107	187	0.81	.000002	2103	187	.00372
	0.91		3105	163	0.80		1748	77	0.93		3085	151	0.93		3165	158	.00118
	1.03		3833	201	1.01		6028	368	1.03		3306	151	1.04		3384	159	.00104
	1.16		3180	167	1.29		5573	419	1.16		3107	160	1.16		3083	160	.00128
	1.35		2373	184	1.64		1412	122	1.33		2259	159	1.34		2186	154	.00234
			954	157			91	9			1136	192			1079	182	.01338

ACCEPTED MANUSCRIPT

Table 3: OSB and Optimum sample sizes for skewness = 1.3466, μ = -0.0132848, σ = 0.4077880, N = 1000, n = 100, nclass= 50, x_0 = 0.28757730 and d = 3.18622440.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	1.12	.000653	633	46	1.00	.000704	514	30	1.19	.000641	689	54	1.17	.000642	682	53
			367	54			486	70			311	46			318	47
3	0.92	.000317	440	35	0.66	.000453	160	6	0.88	.000315	378	27	0.88	.000316	381	28
	1.43		384	29	1.51		692	68	1.36		399	27	1.39		421	31
			176	36			148	26			223	46			198	41
4	0.80	.000187	293	24	0.54	.000299	65	2	0.82	.000183	323	28	0.82	.000183	319	27
	1.12		340	21	1.00		449	32	1.19		366	26	1.17		362	25
	1.63		253	25	1.86		422	52	1.69		216	21	1.66		220	21
			114	30			64	14			95	25			99	27
5	0.73	.000117	229	20	0.47	.000192	35	1	0.73	.000116	227	20	0.75	.000117	244	22
	0.99		274	17	0.78		242	14	0.99		271	16	1.02		287	18
	1.31		256	19	1.28		472	44	1.29		253	18	1.33		235	17
	1.82		165	19	2.11		221	33	1.74		163	17	1.79		154	16
			76	25			30	8			86	29			80	27
6	0.67	.000087	169	15	0.44	.000137	23	1	0.73	.000083	227	24	0.70	.000087	198	19
	0.92		271	20	0.66		137	7	0.97		261	18	0.93		242	16
	1.18		245	18	1.00		354	27	1.22		221	15	1.17		239	17
	1.50		163	15	1.51		338	39	1.55		152	14	1.47		160	14
	1.94		104	13	2.29		130	21	2.05		105	15	1.94		112	15
			48	19			18	5			34	14			49	19

Table 4: OSB and Optimum sample sizes for skewness = 1.7274, μ = -0.004327391, σ = 0.506233130, N = 4000, n = 400, nclass = 50, x_0 = 0.14586200 and d = 6.4382790.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	1.30	.000277	2829	211	0.98	.000318	1934	92	1.28	.000276	2757	199	1.28	.000276	2767	201
			1171	189			2066	308			1243	201			1233	199
3	0.92	.000134	1726	116	0.52	.000262	387	9	0.97	.000133	1910	139	0.97	.000133	1916	140
	1.69		1689	146	1.85		3164	314	1.68		1489	119	1.69		1500	122
			585	138			449	77			601	142			584	138
4	0.79	.000079	1293	91	0.38	.000171	108	2	0.79	.000077	1306	93	0.82	.000078	1416	107
	1.30		1536	118	0.98		1826	106	1.22		1305	82	1.28		1350	92
	2.08		878	98	2.54		1938	262	1.88		964	93	1.99		882	90
			293	93			128	30			425	2132			352	111
5	0.66	.000051	867	59	0.31	.000114	50	1	0.73	.000050	1061	82	0.73	.000050	1083	85
	1.05		1288	90	0.67		836	33	1.08		1175	76	1.08		1170	76
	1.43		904	63	1.43		2174	200	1.51		952	77	1.52		936	75
	2.21		719	100	3.07		891	151	2.22		595	77	2.23		596	77
			222	88			49	15			217	88			215	87
6	0.66	.000037	867	69	0.28	.000085	25	1	0.67	.000035	883	72	0.67	.000036	896	74
	0.92		859	48	0.52		362	12	0.93		898	52	0.96		999	64
	1.30		1103	94	0.98		1547	104	1.22		830	52	1.29		895	65
	1.69		586	49	1.85		1617	194	1.61		691	58	1.72		661	64
	2.46		434	67	3.49		423	80	2.27		495	70	2.46		396	58
			151	73			26	9			203	96			153	75

ACCEPTED MANUSCRIPT

Table 5: OSB and Optimum sample sizes for skewness = 2.1145, $\mu = -0.008319588$, $\sigma = 0.605562077$, $\mu = -0.008319588$, $N = 2000$, $n = 200$, nclass = 50, $x_0 = 0.11417550$ and $d = 6.83472120$.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	1.34	.000892	1387	90	0.89	.001181	855	29	1.39	.000890	1418	95	1.41	.000891	1428	96
			613	110			1145	171			582	105			572	104
3	0.93	.000418	909	54	0.45	.000736	185	3	0.98	.000416	978	62	1.00	.000417	1010	66
	1.89		801	66	1.77		1484	128	1.89		732	58	1.96		727	61
			290	80			331	69			290	80			263	73
4	0.80	.000243	727	46	0.32	.000477	59	1	0.81	.000242	740	48	0.82	.000243	751	49
	1.34		660	43	0.89		796	37	1.36		655	43	1.39		667	45
	2.30		445	48	2.49		1024	128	2.24		424	42	2.37		428	48
			168	63			121	34			181	67			154	58
5	0.66	.000156	515	32	0.26	.000316	29	1	0.72	.000151	591	41	0.73	.000151	608	43
	1.07		592	36	0.59		359	11	1.15		594	38	1.16		583	37
	1.62		471	36	1.34		998	80	1.73		464	40	1.73		458	39
	2.57		315	45	3.06		557	89	2.72		259	36	2.72		259	36
			107	51			57	19			92	45			92	45
6	0.66	.000108	515	39	0.23	.000214	18	1	0.65	.000106	498	37	0.64	.000106	474	34
	1.07		592	43	0.45		167	4	0.98		482	28	0.97		498	30
	1.48		388	28	0.89		670	38	1.37		422	30	1.37		435	32
	2.03		265	25	1.77		814	90	1.90		312	28	1.95		324	32
	2.98		177	29	3.50		291	52	2.90		217	37	2.96		205	35
			63	36			40	15			69	40			64	37

Table 6: OSB and Optimum sample sizes for skewness = 3.5009, $\mu = -0.00328841$, $\sigma = 0.69666629$, $N = 15000$, $n = 1500$, nclass = 50, $x_0 = 0.05041042$ and $d = 22.44861984$.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	1.40	.000197	10293	560	1.06	.000234	8025	295	1.60	.000193	11273	715	1.62	.000193	11353	729
			4707	940			6975	1205			3727	785			3647	771
3	0.95	.000091	7036	354	0.39	.000222	1289	14	1.05	.000090	7938	456	1.09	.000090	8269	497
	2.30		6220	550	2.94		12826	1256	2.29		5318	443	2.40		5156	448
			1744	596			885	230			1744	601			1575	555
4	0.95	.000053	7036	463	0.23	.000137	271	2	0.86	.000051	6283	373	0.87	.000051	6324	378
	1.85		5153	414	1.06		7754	353	1.60		4973	329	1.62		5026	341
	3.64		2346	351	4.90		6803	1065	2.95		2863	343	3.01		2815	345
			465	272			172	80			881	455			835	436
5	0.50	.000037	2438	85	0.17	.000095	69	1	0.72	.000032	4798	280	0.74	.000032	5004	303
	0.95		4598	218	0.58		3203	83	1.22		4378	248	1.28		4552	279
	1.85		5153	478	1.96		9234	837	1.93		3231	254	2.05		3187	277
	3.19		2124	290	6.64		2445	544	3.28		1950	283	3.53		1742	274
			687	429			49	35			643	435			515	367
6	0.50	.000023	2438	105	0.14	.000070	30	1	0.62	.000021	3740	214	0.66	.000022	4135	257
	0.95		4598	269	0.39		1259	22	1.00		3717	187	1.08		4005	230
	1.40		3257	192	1.06		6736	366	1.47		3208	202	1.62		3208	233
	2.30		2963	340	2.94		6090	841	2.17		2349	217	2.43		2124	229
	4.09		1433	311	8.14		860	247	3.50		1455	241	3.98		1186	236
			311	283			25	23			531	439			342	315

ACCEPTED MANUSCRIPT

Table 7: OSB and Optimum sample sizes for skewness = 4.2624, μ = 0.0008461947, σ = 0.8006085337, N = 5000, n = 400, nclass = 50, x_0 = 0.04706870 and d = 26.12477998.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	1.61	.001221	3612	147	1.11	.001510	2787	69	1.89	.001193	3928	191	1.89	.001193	3932	192
			1388	253			2213	331			1072	209			1068	208
3	1.09	.000543	2748	109	0.39	.001139	581	5	1.19	.000536	2959	130	1.20	.000536	2976	132
	2.66		1701	114	3.18		4034	299	2.89		1567	111	2.97		1574	115
			551	177			385	96			474	159			450	153
4	1.09	.000314	2748	143	0.23	.000705	162	1	0.92	.000307	2312	99	0.92	.000307	2306	99
	2.14		1404	85	1.11		2625	88	1.93		1660	98	1.89		1623	92
	4.23		668	76	5.39		2121	272	3.95		817	93	3.86		843	93
			180	96			92	39			211	110			228	116
5	0.57	.000203	1190	35	0.17	.000496	50	1	0.75	.000186	1843	79	0.77	.000187	1882	82
	1.09		1558	57	0.59		1207	23	1.38		1442	65	1.44		1478	71
	2.14		1404	102	2.09		2858	196	2.36		1003	69	2.48		991	73
	4.23		668	91	7.39		851	159	4.35		543	73	4.62		515	78
			180	115			34	21			169	114			134	96
6	0.57	.000131	1190	42	0.13	.000349	21	1	0.67	.000126	1571	70	0.67	.000128	1558	69
	1.09		1558	70	0.39		560	7	1.18		1364	60	1.18		1388	62
	1.61		864	38	1.11		2206	89	1.82		919	50	1.89		982	60
	2.66		837	74	3.18		1828	194	2.80		647	54	3.00		630	59
	4.75		426	71	9.13		372	98	4.82		380	61	5.29		346	63
			125	105			13	11			122	105			96	87

Table 8: OSB and Optimum sample sizes for skewness = 3.8763, μ = 0.004467927, σ = 0.887740363, N = 8000, n = 700, nclass = 50, x_0 = 0.05568601 and d = 28.04155725.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	2.30	.000895	6587	374	1.25	.001106	4787	134	2.10	.000888	6369	333	2.17	.000888	6440	346
			1413	326			3213	566			1631	367			1560	354
3	1.18	.000395	4581	197	0.44	.000739	1438	16	1.27	.000392	4829	225	1.31	.000393	4933	238
	3.42		2757	247	3.53		5942	502	3.40		2501	216	3.58		2462	224
			662	256			620	182			670	259			605	238
4	1.18	.000230	4581	258	0.26	.000453	518	4	0.95	.000215	3807	173	0.98	.000217	3919	184
	2.30		2006	124	1.25		4269	166	2.05		2501	154	2.16		2518	166
	5.10		1142	164	5.93		3023	448	4.32		1296	156	4.77		1238	166
			271	154			190	82			396	217			325	184
5	0.62	.000140	2326	78	0.19	.000308	274	2	0.73	.000135	2901	121	0.80	.000140	3189	148
	1.18		2255	85	0.67		2307	52	1.42		2328	111	1.60		2401	134
	2.30		2006	150	2.33		4043	315	2.49		1533	111	2.92		1491	136
	4.54		1052	148	8.09		1296	286	4.69		903	128	5.78		711	130
			361	239			80	45			335	229			208	152
6	0.62	.000102	2326	95	0.16	.000221	145	1	0.65	.000091	2490	108	0.68	.000095	2627	122
	1.18		2255	104	0.44		1293	21	1.20		2169	99	1.28		2233	112
	2.30		2006	182	1.25		3349	153	1.98		1558	99	2.15		1561	112
	3.98		942	127	3.53		2593	317	3.16		1004	97	3.60		981	115
	7.35		367	96	9.96		581	180	5.42		544	98	6.71		465	117
			104	96			39	28			235	199			133	122

Table 9: OSB and Optimum sample sizes for skewness = 4.3091, $\mu = -0.01518327$, $\sigma = 0.99530377$, $N = 10000$, $n = 500$, nclass = 20, $x_0 = 0.02280605$ and $d = 30.29211804$.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	3.05	.002003	8700	300	0.83	.003618	4353	32	2.47	.001945	8202	238	2.54	.001946	8277	247
			1300	200			5647	468			1798	262			1723	253
3	1.54	.000862	6713	185	0.25	.001686	838	2	1.46	.000860	6550	173	1.47	.000860	6562	174
	4.57		2694	156	2.76		7640	252	4.49		2835	164	4.50		2828	164
			593	159			1522	246			615	163			610	162
4	1.54	.000540	6713	236	0.14	.001110	261	1	1.00	.000466	5081	122	1.05	.000467	5277	134
	3.05		1987	76	0.83		4092	50	2.40		3043	113	2.56		3026	122
	7.60		1105	116	5.02		5160	330	5.72		1508	126	6.07		1379	122
			195	72			487	119			368	139			318	122
5	1.54	.000474	6713	254	0.10	.000709	105	1	0.77	.000298	4054	92	0.85	.000302	4458	113
	3.05		1987	82	0.41		1748	12	1.62		2860	82	1.87		2930	100
	6.08		982	78	1.71		5242	149	3.06		1792	85	3.60		1651	97
	18.20		301	80	7.19		2682	267	6.36		998	101	7.38		746	86
			17	6			223	71			296	140			215	104
6					0.08	.000503	49	1	0.65	.000202	3358	75	0.71	.000209	3666	90
					0.25		789	4	1.32		2823	77	1.44		2819	84
					0.83		3515	55	2.32		1869	74	2.55		1806	80
					2.76		4125	201	3.89		1118	70	4.47		1088	84
					9.14		1383	186	7.22		609	76	8.68		464	72
							139	53			223	128			157	90

Table 10: OSB and Optimum sample sizes for skewness = 5.4744, $\mu = -0.009671336$, $\sigma = 1.104066811$, $N = 5000$, $n = 400$, nclass = 50, $x_0 = 0.01439434$ and $d = 46.00158162$.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	2.77	.003678	4130	169	0.81	.007997	2129	19	3.17	.003627	4295	201	3.17	.003627	4295	201
			870	231			2871	381			705	199			705	199
3	1.85	.001535	3559	152	0.21	.003152	422	1	1.54	.001501	3256	117	1.69	.001518	3420	136
	5.53		1146	100	3.12		3859	185	5.36		1437	129	5.86		1311	126
			295	148			719	214			306	154			269	138
4	0.93	.000800	2362	68	0.11	.002030	113	1	1.08	.000785	2641	89	1.17	.000795	2798	104
	2.77		1768	106	0.81		2016	31	2.77		1486	82	3.16		1492	99
	7.37		699	101	6.12		2625	255	7.07		689	96	8.23		574	93
			171	125			246	113			184	133			136	104
5	0.93	.000476	2362	84	0.07	.001318	48	1	0.80	.000465	2114	64	0.91	.000500	2321	81
	1.85		1197	46	0.36		876	7	1.79		1397	57	2.17		1504	80
	3.69		867	65	1.82		2616	103	3.56		885	63	4.55		757	74
	8.29		438	79	9.16		1344	216	7.67		442	70	10.32		318	68
			136	126			116	73			162	146			100	97
6	0.93	.000362	2362	100	0.06	.000942	26	1	0.64	.000297	1725	50	0.75	.000343	1991	71
	1.85		1197	56	0.21		396	2	1.35		1304	46	1.63		1364	63
	2.77		571	27	0.81		1707	33	2.39		923	48	3.04		896	67
	5.53		575	78	3.12		2152	153	4.19		572	51	5.64		457	59
	11.05		214	58	11.99		657	161	8.11		337	66	12.18		230	78
			81	81			62	50			139	139			62	62

ACCEPTED MANUSCRIPT

Table 11: OSB and Optimum sample sizes for skewness = 6.6147, μ = −0.01056465, σ = 1.2029671, N = 3000, n = 150, nclass = 50, x_0 = 0.02222465 and d = 65.25438173.

L	Cum $\sqrt{f}$				Geometric				L-H (Kozak Algo.)				Dynamic Prog.			
	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h	OSB	$V(\bar{x}_{st})^*$	N_h	n_h
2	3.94	.014646	2608	76	1.20	.078002	1701	14	3.55	.014453	2562	70	3.88	.014520	2602	76
			392	74			1299	136			438	80			398	74
3	1.33	.006523	1786	29	0.32	.011590	519	2	1.77	.006265	2043	44	1.95	.006361	2143	51
	5.24		963	45	4.56		2180	81	56.60		784	47	7.59		722	50
			251	76			301	67			173	59			135	49
4	1.33	.003452	1786	39	0.16	.007350	204	1	1.25	.003393	1731	36	1.32	.003483	1778	38
	3.94		822	35	1.20		1497	19	3.50		825	32	3.87		823	35
	9.16		295	26	8.87		1199	94	9.23		347	32	10.94		333	39
			97	50			100	36			97	50			66	38
5	1.33	.002287	1786	46	0.11	.004999	108	1	0.91	.002068	1387	25	0.99	.002089	1506	31
	2.63		580	15	0.54		802	5	2.35		903	28	2.51		848	29
	5.24		383	21	2.68		1470	45	5.20		458	28	5.58		428	29
	10.46		176	18	13.22		584	79	11.68		196	27	13.95		187	33
			75	50			36	20			56	42			31	28
6	1.33	.001741	1786	52	0.08	.003261	66	1	0.76	.001344	1235	23	0.81	.001389	1298	27
	2.63		580	18	0.32		453	2	1.75		800	21	1.90		811	25
	3.94		242	7	1.20		1182	19	3.48		519	25	3.77		484	25
	6.55		216	14	4.56		998	60	6.75		277	24	7.40		264	27
	11.77		120	14	17.25		282	53	14.18		140	28	16.64		122	25
			56	45			19	15			29	29			21	21